Riley R, Haridas S, Wolfe KH, Lopes MR, Hittinger CT, Göker M, Salamov A, Wisecaver J, Long TM, Aerts AL, Barry K, Choi C, Clum A, Coughlan AY, Deshpande S, Douglass AP, Hanson SJ, Klenk HP, LaButti K, Lapidus A, Lindquist E, Lipzen A, Meier-Kolthoff J, Ohm RA, Otillar RP, Pangilian J, Peng Y, Rokas A, Rosa CA, Scheuner C, Sibirny A, Slot JC, Stielow B, Sun H, Kurtzman CP, Blackwell M, Grigoriev IV, Jeffries TW.

Comparative genomics of biotechnologically important yeasts.

***Proceedings of the National Academy of Sciences of the USA* 2016, 113(35), 9882-9887.**

DOI link to article:

http://dx.doi.org/10.1073/pnas.1603941113

Date deposited:

14/10/2016

Embargo release date:

17 February 2017

Comparative genomics of biotechnologically important yeasts

Robert Riley[1], Sajeet Haridas[1], Kenneth H. Wolfe[2], Mariana R. Lopes[3,4], Chris Todd Hittinger[3,5], Markus Göker[6]; Asaf Salamov[1], Jen Wisecaver[7], Tanya M. Long[8], Andrea L. Aerts[1], Kerrie Barry[1], Cindy Choi[1], Alicia Clum[1], Aisling Y. Coughlan[2], Shweta Deshpande[1], Alexander P. Douglass[2], Sara J. Hanson[2], Hans-Peter Klenk[6,9], Kurt LaButti[1], Alla Lapidus[1], Erika Lindquist[1], Anna Lipzen[1], Jan Meier-Kolthoff[6], Robin A. Ohm[1], Robert P. Otillar[1], Jasmyn Pangilinan[1], Yi Peng[1], Antonis Rokas[7], Carlos A. Rosa[4], Carmen Scheuner[6], Andryi Sibirny[10], Jason C. Slot[11], Benjamin Stielow[6,12], Hui Sun[1], Cletus P. Kurtzman[13], Meredith Blackwell[14,15], Igor V. Grigoriev[1], Thomas W. Jeffries[8]

[1]Department of Energy Joint Genome Institute, Walnut Creek, CA 94598; [2]UCD Conway Institute, School of Medicine, University College Dublin, Dublin 4, Ireland; [3]Laboratory of Genetics, Genome Center of Wisconsin, Wisconsin Energy Institute, J. F. Crow Institute for the Study of Evolution, University of Wisconsin-Madison, Madison, WI; [4]Departamento de Microbiologia, ICB, C.P. 486, Universidade Federal de Minas Gerais, Belo Horizonte, MG, 31270-901, Brazil; [5]DOE Great Lakes Bioenergy Research Center, University of Wisconsin-Madison, Madison, WI; [6]Leibniz Institute DSMZ – German Collection of Microorganisms and Cell Cultures, Inhoffenstraße 7B, 38124 Braunschweig, Germany; [7]Department of Biological Sciences, Vanderbilt University, Nashville, TN 37235, USA; [8]University of Wisconsin-Madison, Department of Bacteriology and USDA Forest Products Laboratory, Madison, WI, USA; [9]Newcastle University, School of Biology, Ridley Building, Newcastle upon Tyne, UK; [10]Department of Molecular Genetics and Biotechnology, Institute of Cell Biology, NAS of Ukraine, Lviv 79005 Ukraine; [11]College of Food, Agricultural, and Environmental Sciences, Ohio State University, Columbus, OH; [12]CBS-KNAW Fungal Biodiversity Centre, Utrecht, the Netherlands; [13]US Department of Agriculture ARS NCAUR 1815 N University St. Peoria IL 61604; [14]Department of Biological Sciences, Louisiana State University, Baton Rouge, LA 70803 USA; [15]Department of Biological Sciences, University of South Carolina, Columbia, SC, 29208 USA

Present addresses: Center for Algorithmic Biotechnology, St. Petersburg State University, Russia (Alla Lapidus); Microbiology, Department of Biology, Utrecht University, Utrecht, The Netherlands (Robin A. Ohm)

C.P.K., M.B., I.V.G., and T.W.J. designed the study. K.B. C.P.K., M.B., I.V.G. and T.W.J. coordinated the project. T.M.L., C.C., A.Y.C, S.D., S.J.H, H.K., Y.P., A.S., B.S., C.P.K. and T.W.J. performed experiments. R.R., S.H., K.H.W., M.R.L., C.T.H., M.G., A. Salamov, J.W., A.L.A., A.C., A.P.D., K.L., A. Lapidus, E.L., A. Lipzen, J.M., R.A.O., R.P.O., J.P., Y.P., A.R., C.A.R, C.S., J.C.S., C.P.K., I.V.G. and T.W.J. analyzed data. R.R., S.H., K.H.W., C.T.H., M.G., C.P.K., M.B., I.V.G. and T.W.J. wrote the manuscript.

Abstract

Ascomycete yeasts have high metabolic diversity and great potential for biotechnology. Here analyzed genomes of 38 taxonomically and biotechnologically important species, including 16 new sequences. We identify a genetic code change, CUG-Ala, in *Pachysolen tannophilus* which is sister to the known CUG-Ser clade. Our well-resolved yeast phylogeny shows that some traits such as methylotrophy are restricted to single clades, whereas others such as L-rhamnose utilization have patchy phylogenetic distributions. Many pathways of interest are encoded by gene clusters, with variable organization and distribution. Genomics can predict some biochemical traits precisely, but the genomic basis of others such as xylose utilization remains unresolved. Our data also provides insight into early evolution of ascomycetes. We document the loss of H3K9me2/3 heterochromatin, the origin of ascomycete mating-type switching, and pan-ascomycete synteny at the MAT locus. These data and analyses provide strategies for engineering efficient biosynthetic and degradative pathways, and gateways for genomic manipulation.

Introduction

Yeasts are fungi that reproduce asexually, by budding or fission, and sexually without multicellular fruiting bodies[1, 2]. Their unicellular, largely free-living lifestyle has evolved several times in the fungi[3]. Despite morphological similarities, yeasts constitute approximately 1,500 known species that inhabit many specialized environmental niches and associations including virtually all varieties of fruits and flowers, plant surfaces and exudates, insects, mammals and highly diverse soils[4]. Biochemical, molecular biological, and genomic studies of the model yeast *Saccharomyces cerevisiae* - essential for making bread, beer, and wine - have established much of our understanding of eukaryotic biology. However, in many ways, *S. cerevisiae* is an oddity among the yeasts, and many important biotechnological applications and highly divergent physiological capabilities of lesser-known yeast species have not been fully exploited[5]. Various species can grow on minimal media, use methanol, produce vitamins, accumulate lipids, thrive under acidic conditions and ferment unconventional carbon sources. Many features of yeasts make them ideal platforms for biotechnological processes. Their thick cell walls help them survive osmotic shock, and they are less susceptible to viruses than bacteria. Their unicellular form is easy to cultivate, scale up and harvest. The objective of this study was therefore to put yeasts with diverse biotechnological applications in a phylogenomic context, and to relate their physiologies to genomic features so that their useful properties may be developed through genetic techniques. Backgrounds on the 16 yeasts (Table 1) and detailed justifications for their choice are given in Supplementary Note 1.

Organism	**Genome size (Mb)**	**Number of genes predicted**
Ascoidea rubescens NRRL Y-17699	17.5	6,802
Babjeviella inositovora NRRL Y-12698	15.2	6,403
Candida arabinofermentans NRRL YB-2248	13.2	5,861
Candida tanzawaensis NRRL Y-17324	13.1	5,895
Cyberlindnera jadinii NRRL Y-1542	13.0	6,038
Hanseniaspora valbyensis NRRL Y-1626	11.5	4,800
Hyphopichia burtonii NRRL Y-1933	12.4	6,002
Lipomyces starkeyi NRRL Y-11557	21.3	8,192
Metschnikowia bicuspidata NRRL YB-4993	16.1	5,851
Nadsonia fulvescens var. *elongata* DSM 6958	13.7	5,657
Ogataea polymorpha NCYC 495 leu1.1	9.0	5,177

Pachysolen tannophilus NRRL Y-2460	12.6	5,675
Pichia membranifaciens NRRL Y-2026	11.6	5,546
Saitoella complicata NRRL Y-17804	14.1	7,034
Tortispora caseinolytica NRRL Y-17796	9.2	4,657
Wickerhamomyces anomalus NRRL Y-366-8	14.1	6,423

Table 1. Yeasts chosen for sequencing. Genomes and annotations are accessible at http://jgi.doe.gov/fungi. All are from the subphylum Saccharomycotina, except *Saitoella complicata* (Taphrinomycotina). Genome features and sequencing statistics are given in Supplementary Note 2, Supplementary Tables 1 and 2, and Supplementary Figures 1-4.

Results

Organism phylogeny

Using the entire proteome sequences of 30 yeast species and eight outgroups, we generated three phylogenomic data matrices: 'full' (7,297 genes with ≥4 sequences), 'MARE' (1,559 genes from the 'full' set filtered for informative quality), and 'core' (418 genes present in all organisms). The MARE-filtered supermatrix tree[6, 7] is shown (Fig. 1). The full-supermatrix and core gene trees differed regarding the position of *Ascoidea rubescens*, *Debaryomyces hansenii* and *Metschnikowia bicuspidata*, but conflicting branches were not supported. The MP MARE-filtered supermatrix and core-genes matrix trees were topologically identical to Figure 1 (Supplementary Note 3; Supplementary Figs. 5 and 6). Overall, our data show a phylogenetic tree with three major Saccharomycotina clades (CUG-Ser, Methylotrophs clade, and Saccharomycetaceae) along with significant divergence of the early diverging members such as *Lipomyces starkeyi, Tortispora caseinolytica, Yarrowia lipolytica,* and *Nadsonia fulvescens*.

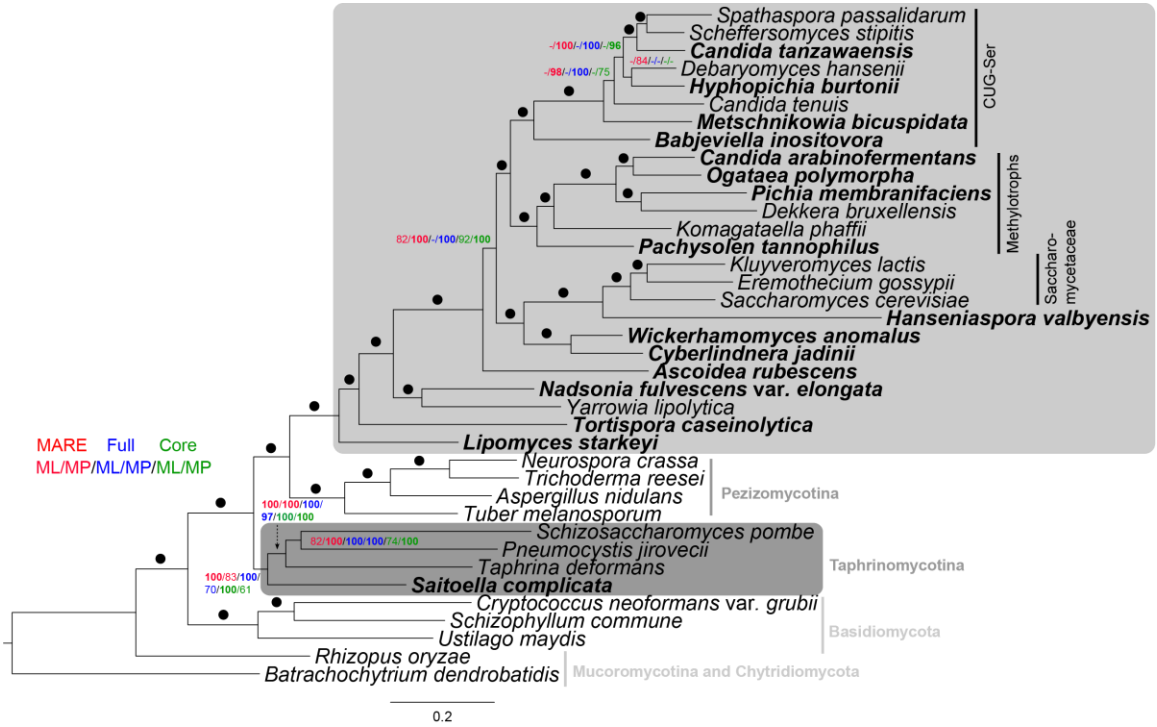
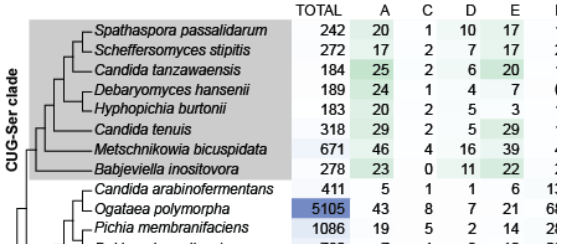


Figure 1. Phylogenetic tree inferred from the MARE-filtered supermatrix (364,126 aligned amino acid residues) using maximum likelihood (ML) and rooted with *Batrachochytrium*. Organisms sequenced in this study are named in bold. Numbers on the branches indicate ML and maximum-parsimony (MP) bootstrap support values for the MARE-filtered (red), full (blue) and core-genes (green) supermatrices. Values less than 60% are shown as '-'; dots indicate branches with maximum support under all settings.

Alternative genetic codes: CUG coding for Ser and Ala

Yeasts in the CUG-Ser clade, including *Candida albicans*, use an altered genetic code in which CUG codons are translated as Ser rather than the canonical Leu[8-13] due to alterations in the $tRNA_{CAG}$ that decodes CUG. To investigate the origins of this change, we inspected predicted $tRNA_{CAG}$s for the presence of three sequence features indicative of Ser translation[14]: a G33 residue 5' to the anticodon, which may lower rates of leucylation[15], a Ser identity element in the variable loop, and a G discriminator base. Most CUG-Ser clade species contained all three serylation features in their predicted $tRNA_{CAG}$s, indicating translation of CUG to Ser (Supplementary Fig. 7). However in the most basal taxa of this clade not all of the features are present: *Metschnikowia bicuspidata* lacks the Ser identity element, and *Babjeviella inositovora* lacks the discriminator base. This may reflect stepwise accumulation of $tRNA_{CAG}^{Ser}$ features in the evolution of alternative CUG translation. Species branching deeper in the tree do not show any of the three features.

To investigate CUG translation in a broader phylogenetic context, we analyzed multiple alignments of 700 orthologous groups of proteins (Supplementary Note 4). For each yeast we identified its CUG-encoded positions in the alignments and tabulated the frequencies of the amino acids in other species to which CUG sites aligned, restricting the analysis to conserved regions of proteins. In the CUG-Ser clade, CUG codons most frequently aligned with Ser rather than Leu (Fig. 2).



	TOTAL	A	C	D	E
Spathaspora passalidarum	242	20	1	10	17
Scheffersomyces stipitis	272	17	2	7	17
Candida tanzawaensis	184	25	2	6	20
Debaryomyces hansenii	189	24	1	4	7
Hyphopichia burtonii	183	20	2	5	3
Candida tenuis	318	29	2	5	29
Metschnikowia bicuspidata	671	46	4	16	39
Babjeviella inositovora	278	23	0	11	22
Candida arabinofermentans	411	5	1	1	6
Ogataea polymorpha	5105	43	8	7	21
Pichia membranifaciens	1086	19	5	2	14

Figure 2. Amino acids aligned to yeast CUG codons based on 700 orthologous groups of proteins

In the eight genomes from this clade, CUG codons aligned with Ser in 32–56% of aligned positions. For most of the other yeasts CUG aligned predominantly with Leu (70–86%). However, two yeasts outside the CUG-Ser clade, *Pachysolen tannophilus* and *Ascoidea rubescens*, show unusual CUG alignment patterns. They were previously proposed to translate CUG as Ser on the basis of interspecies alignments[16, 17] (Supplementary Table 3), but their $tRNA_{CAG}$ genes lack the serylation features (Supplementary Fig. 7). In *P. tannophilus*, CUG unexpectedly aligned mostly with Ala (29%; Fig. 2), more than with Leu (7%) or Ser (8%). In *A. rubescens*, CUG codons are remarkably rare, and showed no strong preference for any amino acid.

To determine the genetic code of *P. tannophilus* directly we sequenced tryptic peptides *de novo* by LC-MS/MS, and compared them to the genome sequence. Among 6,836 peptides that mapped to unique sites in the genome, 178 span a CUG codon site (in 170 different genes). Of these CUG codon sites, 160 (90%) align with Ala in the sequenced peptide, 16 with Leu, and 2 with other amino acids (Supplementary Table 4). The possibility of mRNA editing can be excluded because no editing was seen at these sites in expressed sequence tag (EST) data from *P. tannophilus*, for all 166 sites that were covered by EST reads. We conclude that *P. tannophilus* translates most CUG codons in mRNA as Ala.

Correlation of genomically encoded enzymes to metabolic traits

To determine how traits are genetically controlled in these diverse yeasts, we correlated several metabolic capabilities[2] with genome content (Fig. 3; Supplementary Note 5; Supplementary Figs. 8 and 9). Some traits of biotechnological interest, such as D-xylose assimilation, did not correlate well with genetically characterized pathways (Fig. 3). Profile-based searches[18] of predicted proteins yielded homologs of xylose reductase (Xyl1), xylitol dehydrogenase (Xyl2), and D-xylulokinase (Xyl3) in the CUG-Ser clade in which xylose metabolism has been well-characterized[19, 20], but Xyl1 and Xyl2 were not found in more phylogenetically distant yeasts, suggesting sequence divergence of bioactive enzymes. Unexpectedly, we found *Candida tanzawaensis* did not grow on xylose despite possessing predicted homologs for the full pathway, suggesting that, rather than simple gene gain or loss, variation in this trait may involve changes in enzyme kinetics, gene regulation, cofactor balancing, regulation of glycolytic flux, sugar transport or respiration efficiency. Conversely, many yeasts beyond the CUG-Ser clade are able to metabolize D-xylose despite lacking predicted Xyl1 and/or Xyl2 genes, suggesting more distant homologs may perform these functions.

Galactose utilization. In contrast to D-xylose, the utilization of galactose as a carbon source correlates well with thoroughly characterized pathways from *S. cerevisiae*[21]. The ability to utilize galactose varies widely across yeasts[2], and previous research has documented a handful of losses of genes in the galactose (*GAL*) utilization pathway and at least one re-acquisition by horizontal gene transfer[22-27]. The denser sampling of yeast genomes in this study provides a much richer picture of the dynamic evolution of this pathway. Parsimony suggests that key *GAL* genes and the ability to utilize galactose have been lost at least seven times among the examined taxa, for a total of at least 11 known losses among the subphyla Saccharomycotina and Taphrinomycotina (Fig. 3). Since *GAL* homologs are conserved across all domains of life[28], our analyses suggest that the dominant mode of evolution for yeast galactose consumption is one of repeated and independent loss of the trait, along with its required genes, from an ancestral yeast that could consume galactose.

Methylotrophy. Relatively few yeasts can use methanol as a sole carbon source[29]. All three yeasts in this study that use methanol - *Ogataea polymorpha*, *Candida arabinofermentans* and *Komagataella phaffii* - were found to possess a full complement of genes for methanol utilization pathway enzymes (Fig. 3) and belong to the same multigenus clade (Fig. 1). In particular this rare trait correlates perfectly with the presence of genes encoding alcohol oxidases (AOX) and dihydroxyacetone synthases (DAS) but not with other members of the pathway. Methylotrophy appears to have been lost, due to loss of AOX and DAS, in *P. membranifaciens* and *D. bruxellensis*, which likely had a methylotrophic ancestor.

L-Rhamnose utilization. L-Rhamnose metabolism can proceed by phosphorylated (isomerase) and non-phosphorylated (oxidative) pathways with the latter occurring in fungi[30]. Its catabolism requires several enzymatic steps, all of which are necessary to enter central metabolism. In *S. stipitis*, the four genes of the oxidative L-rhamnose pathway are situated side-by-side in a cluster, which is repeated in various conformations among six of the yeasts in this study: *C.*

tenuis, *D. hansenii*, *O. polymorpha*, *L. starkeyi* and *K. phaffii*, in addition to *S. stipitis,* and the occurrence of the conserved cluster across broad phylogenetic distances correlates with a yeast's capacity to oxidize L-rhamnose. However L-rhamnose assimilation is sparsely distributed throughout Saccharomycotina (Fig. 3), with losses of the pathway apparent in several clades

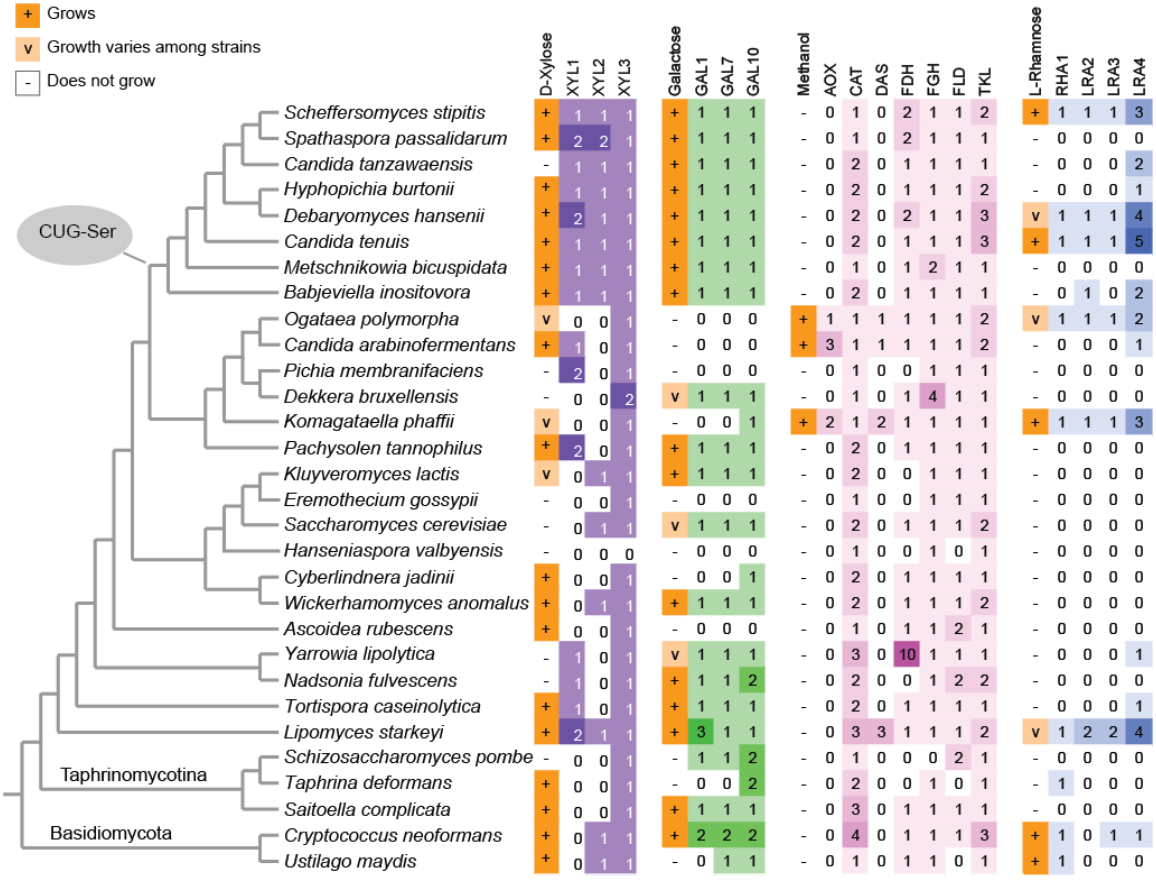


Figure 3. Distribution of metabolic traits and their genes. For D-xylose, galactose, L-rhamnose and methanol, a "+" indicates utilization and boxes indicate the numbers predicted pathway genes. Genes for D-xylose metabolism were annotated with PRIAM; genes for galactose, methanol, and L-rhamnose metabolism were identified through BLAST homology to characterized sequences (see Methods).

Complex I (NADH Dehydrogenase) / DHODase (URA9/URA1)

Our broad phylogenetic study confirmed and elaborated prior findings[31] that loss of respiration Complex I (RC1) preceded the gain of bacterial dihydroorotate dehydrogenase (*URA1*) in *S. cerevisiae* and closely related genera. These two features along with other evolutionary changes such as expansion of genes for facilitated uptake of sugars enabled anaerobic growth in these yeasts. Unexpectedly, two other fermentative species (*Nadsonia* and *Schizosaccharomyces*) showed diminished RC1 components, but did not acquire *URA1* (Supplementary Note 6; Supplementary Fig. 10)

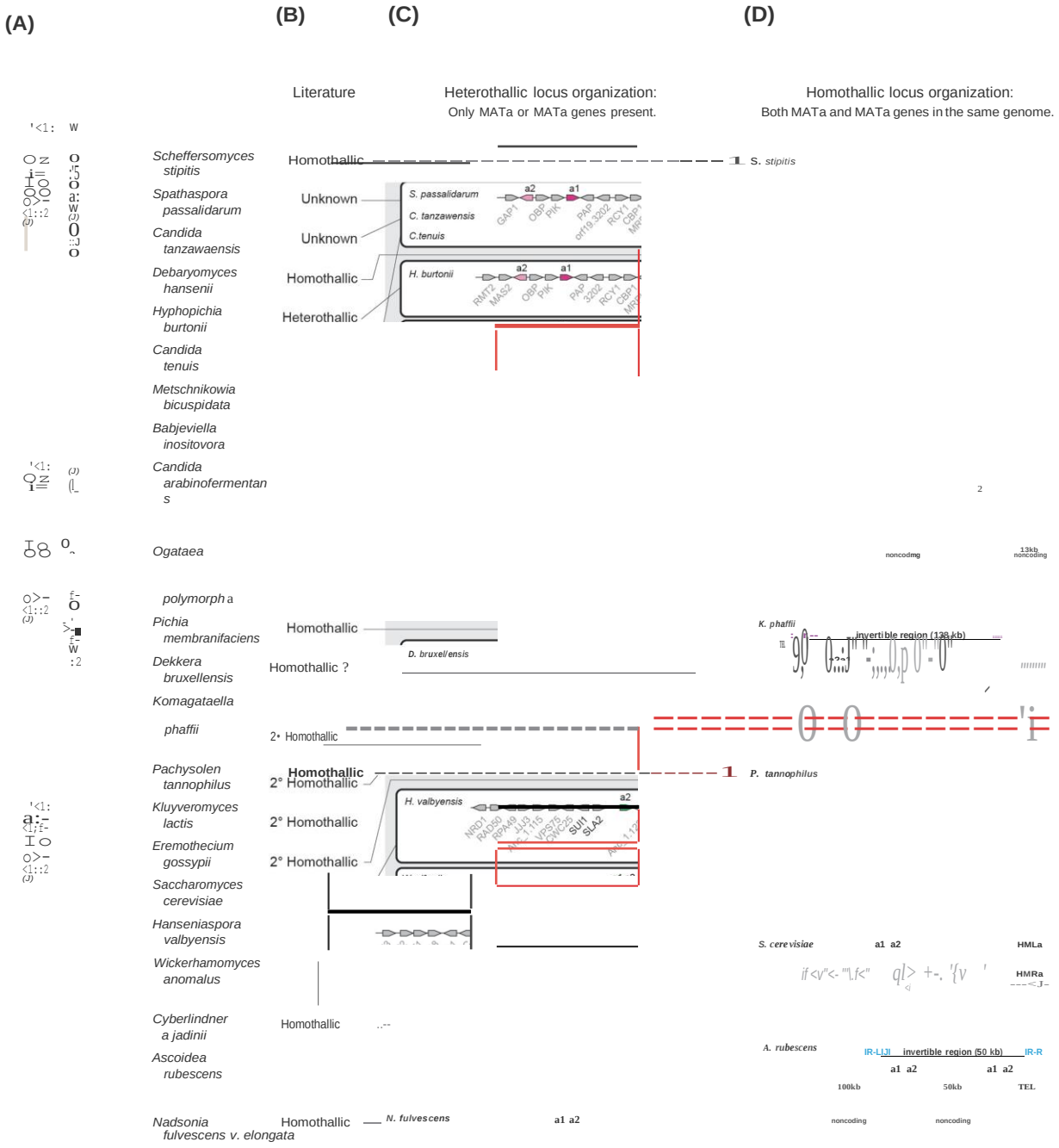
Metabolic gene clusters

Many genes for degradative and biosynthetic traits were proximal on chromosomes across wide phylogenetic ranges (Supplementary Note 7; Supplementary Table 5). Our analysis confirmed previously recognized gene clusters for urea, allantoin, galactose and N-acetyl glucosamine catabolism, starch, cellulose and nitrate utilization, and extended these associations across

multiple genomes. Genes for lipid synthesis and amino acid metabolism (Gly, Ser, Phe, Tyr, Trp) likewise were found in clusters. Three successive enzymatic steps in the biotin synthesis pathway were found in pairwise clusters that differed with phylogeny. Genes for the first two of these steps are frequently clustered in Saccharomycotina [32], whereas clusters with genes for the second and third steps were prevalent in the other clades.

MAT locus organization, mating-type switching, and H3K9me heterochromatin

Mating-type (*MAT*) locus structures of the sequenced species (Fig. 4) support previous biological data about homothallism or heterothallism of the species. Comparison of *MAT* loci revealed, for the first time, evidence of conservation of synteny at this locus among all three subphyla of Ascomycota – showing that cell type has been controlled during one billion years of ascomycete evolution by a single orthologous locus (*MAT*), despite gross changes of its gene content (Supplementary Note 8). The stability of *MAT* contrasts with the frequent turnover of sex-determination loci in animals and plants. We also found that *P. tannophilus* and *A. rubescens* have mating-type switching systems that operate by inversion of a region of chromosome, similar to *O. polymorpha* and *K. phaffii* [33, 34] (Fig. 4; Supplementary Fig. 10). Most Saccharomycotina species have lost the ancestral form of eukaryotic heterochromatin in which lysine 9 of histone H3 is methylated[35]; *Lipomyces starkeyi* is the only Saccharomycotina species with orthologs of *Schizosaccharomyces pombe* Clr4 (H3K9 methyltransferase), Epe1 (H3K9 demethylase) and Swi6 (H3K9me2/3-binding chromodomain protein). We infer that mating-type switching, which requires a mechanism to silence the non-expressed copies of *MAT* genes, evolved in the Saccharomycotina lineage relatively soon after the ancestral H3K9me2/3 form of heterochromatin was lost (Fig. 4).


(A)
(B)
(C)
(D)
Literature
Heterothallic locus organization:
Only MATa or MATα genes present.
Homothallic locus organization:
Both MATa and MATα genes in the same genome.
Scheffersomyces stipitis
Spathaspora passalidarum
Candida tanzawaensis
Debaryomyces hansenii
Hyphopichia burtonii
Candida tenuis
Metschnikowia bicuspidata
Babjeviella inositovora
Candida arabinofermentans
Ogataea polymorpha
Pichia membranifaciens
Dekkera bruxellensis
Komagataella phaffii
Pachysolen tannophilus
Kluyveromyces lactis
Eremothecium gossypii
Saccharomyces cerevisiae
Hanseniaspora valbyensis
Wickerhamomyces anomalus
Cyberlindnera jadinii
Ascoidea rubescens
Nadsonia fulvescens v. elongata
Yarrowia lipolytica
Tortispora caseinolytica
Lipomyces starkeyi
Trichoderma reesei
Neurospora crassa
Aspergillus nidulans
Tuber
Homothallic
Heterothallic
Unknown
Homothallic ?
Heterothallic:
S. stipitis
S. passalidarum
C. tanzawensis
C.tenuis
H. burtonii
D. bruxellensis
P. tannophilus
H. valbyensis
K. phaffii
S. cerevisiae
A. rubescens
N. fulvescens
Y.lipolytica
T. caseinolytica
L. starkeyi
a1
a2
HMLa
HMRa
invertible region (138 kb)
invertible region (50 kb)
IR-L
IR-R
100kb
50kb
TEL
noncoding
13kb
Het
ero
thal
lic

0 + "

N. crassa

Unknown

<J a1-like homeodomains

<J a2-like homeodomains

<J Pi-like homeodomains

... a2-like HMG domains

. . . a1-like HMG domains

<J Pe-like HMG domains

... Me-like HMG domains

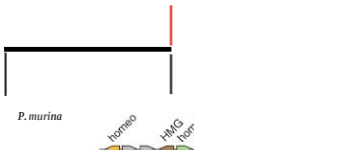


Figure 4. Mating-type locus organization and synteny in Ascomycota. (**A**) Phylogenetic tree showing inferred points of origin of mating-type switching, loss of H3K9me-mediated heterochromatin, and gain of Sir1-mediated silencing and HO endonuclease. (**B**) Summary of literature on homo- or heterothallism of each species[2]. Discrepancies with genome data are highlighted in bold. (**C,D**) Schematic gene organization around *MAT* in selected species. Species were categorized as either heterothallic (C) or homothallic (D) depending on whether *MAT*a and *MAT*alpha genes are present in the same haploid genome assembly. Red outlines indicate species known or predicted to be capable of mating-type switching (secondary homothallism). Gene names in bold (*NVJ2*, *APC5*, *SLA2* and *SUI1*) show conserved synteny among Taphrinomycotina, Pezizomycotina and Saccharomycotina near the *MAT* locus. To maximize data, for *Wickerhamomyces* and *Pneumocystis* we used different congeneric species for the *MAT* locus and phylogenomic analyses. (141 words)

Discussion

By filling in key taxonomic gaps in yeast phylogeny, we have been able to identify major evolutionary transitions in yeast metabolism. For example, pentose and cellobiose fermenting yeasts are often associated with wood-ingesting beetles and are mainly found in the CUG-Ser clade. *P. tannophilus,* which also ferments xylose to ethanol, is in a newly recognized sister CTG-Ala clade. The ethanologenic yeasts, which include *S. cerevisiae, K. lactis,* and several other genera gained the capacity for anaerobic growth by acquiring genes that enable anaerobic uracil synthesis, while losing NADH dehydrogenase. Loss of Respiration Complex I occurred more than once (Supplementary Fig. 9), and diminished respiration correlates with increased fermentative activity. The methylotrophic yeasts occur in a single clade that has retained high oxidative pentose phosphate pathway activity, while gaining tolerance to salt and low pH. The most lipogenic yeasts are closest to the base of divergence from filamentous ascomycetes.

The tree topology presented here (Fig. 1) generally agrees with previous work[1, 36, 37], but almost all branches receive stronger support. The clade including *S. cerevisiae* and *K. lactis* remains strongly supported as the sister of the clade comprising *Wickerhamomyces* and *Cyberlindnera*. Furthermore, there is now strong support for placement of *Dekkera*/*Brettanomyces* in the *Pichia* clade, and for a sister relationship between methylotrophs and the CUG-Ser clade. Inclusion of *Saitoella complicata* in the analysis greatly improved support for a monophyletic grouping of some members of the Taphrinomycotina. Our analysis provides strong support for a number of previous phylogenetic conclusions[1, 36, 37], but more data are needed to resolve the remaining poorly sampled clades. In addition to providing an understanding of biochemical pathway evolution, a strongly supported tree with more inclusive sampling will give a more stable higher-level taxonomy of the yeasts.

Common morphological features seldom resolve phylogenetic classification of yeasts. For example, ascospore shape and presence or absence of pseudohyphae and true hyphae are seldom exclusive to one clade. Although *Eremothecium* species have unique elongated ascospores with a tail-like extension, other morphologies such as hat-shaped ascospores are found in many clades, including members of Pezizomycotina. Bipolar budding is known for four genera, *Hanseniaspora*, *Saccharomycodes*, *Nadsonia* and *Wickerhamia*, but *Hanseniaspora* and *Nadsonia* are only distantly related (Fig. 1). Similarly, many metabolic traits are shared too broadly to have phylogenetic value, such as the fermentation of glucose. However, metabolism of some compounds such as methanol is monophyletic. Species that assimilate methanol belong to the multigenus ‘methylotrophs’ clade (Fig. 1) that includes *Ogataea*, *Kuraishia*, *Komagataella*, *Pichia* and *Dekkera*, but the latter two genera do not metabolize methanol, which suggests loss of this physiological trait in these two lineages.

Certain enzymes are found in some clades to a much greater extent than in others. For example, genes for cellulose utilization are found largely in the CUG-Ser clade where beta-glucosidases, endo-glucanases and transporters occur in functional clusters. *S. stipitis*, a member of the CUG-Ser clade, has six beta-glucosidases, three endoglucanases, and multiple cellobiose transporters along with some xylanase activities. Most of the beta- and endo-glucosidases are induced in response to growth on cellobiose where they confer the ability for rapid assimilation and fermentation of cellobiose. Enzymes for L-rhamnose metabolism are also found in functional clusters in the CUG-Ser yeasts[30, 38]. Based on clusters of the enzymes for L-rhamnose metabolism, it is possible to predict the capacity for L-rhamnose metabolism in other yeasts.

By sampling early-branching lineages of Saccharomycotina, our study elucidates many aspects of genome evolution in this subphylum. Synteny of the MAT locus has been conserved across Ascomycota, and the capacity for sexual recombination has surely contributed to yeasts' metabolic and physiological versatility. Contrastingly, ribosomal DNA organization is highly dynamic (Supplementary Fig. S4). Gene clusters appear to be common across yeast taxa, with the most conspicuous examples attributable to the utilization of substrates or biosynthetic pathways that require several metabolic steps. Reassignment of the CUG codon was more complex than previously thought, affecting a broader phylogenetic range of species, and involved at least two distinct events.

The genetic code change in *P. tannophilus* has practical implications for the use of this species in biotechnology, because heterologous genes containing CUG codons may not produce functional proteins. Indeed, we and others[39] have been unable to transform *P. tannophilus* with the kanamycin (Kan^R) resistance marker gene from Tn903 which contains four CUG codons, whereas the *S. cerevisiae HXK2* gene, which has no CUG codons, was used successfully to complement a *P. tannophilus* mutant[40]. Our discovery that CUG codes for Ala in *P. tannophilus* nuclear genes is only the second known example of a naturally occurring sense codon reassignment in any organism, the other being the CUG to Ser reassignment[41, 42]. All other genetic code changes involve the capture or reassignment of stop codons.

The use of 'non-conventional' yeast species in biotechnology is still in its infancy, and the industry utilizes only a tiny fraction of the thousands of species that are potentially exploitable. The species in widespread use have generally been chosen on the basis of classical assays of their enzymatic or physiological properties in laboratory conditions, without regard to the possible full potential of their genomes. *Sch. pombe* provides an informative illustration. Traditional tests indicate that *Sch. pombe* cannot grow on galactose[2], but its genome contains a set of *GAL* pathway genes predicted to be functional. Indeed, mutants of *Sch. pombe* have been isolated that grow on galactose and constitutively express their *GAL* pathways[44], suggesting that *Sch. pombe* may respond to a different induction signal. Likewise, the *C. tanzawensis* D-xylose pathway may have become rewired. If this type of situation is commonplace, traditional biochemical assays may have overlooked a significant portion of genomic potential. As sequencing costs decrease, mining the genomes of the thousands of currently unsequenced yeast species offers an efficient route towards discovering the next generation of workhorse yeasts for biotechnology.

Acknowledgements

The work conducted by the U.S. Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, is supported by the Office of Science of the U.S. Department of Energy

under Contract No. DE-AC02-05CH11231. TWJ gratefully acknowledges the DOE Great Lakes Bioenergy Research Center (DOE Office of Science BER DE-FC02-07ER64494) and the USDA, Forest Products Laboratory for financial. We thank Marco A. Soares for computational advice. MRL gratefully acknowledges a fellowship from the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES - Brazil, process number 7371/13-6). CAR gratefully acknowledges support from Conselho Nacional de Desenvolvimento Científico e Tecnológico-CNPq. This material is based upon work supported by the National Science Foundation under Grant No. DEB-1442148 to CTH and CPK and funded in part by the DOE Great Lakes Bioenergy Research Center (DOE Office of Science BER DE-FC02- 07ER64494). CTH is a Pew Scholar in the Biomedical Sciences and an Alfred Toepfer Faculty Fellow, supported by the Pew Charitable Trusts and the Alexander von Humboldt Foundation, respectively. MB thanks Drs. S.O. Suh, H. Urbina and N.H. Nguyen and numerous LSU undergraduates for their assistance. KHW thanks G. Cagney, E. Dillon and K. Wynne (UCD Conway Institute Proteomics core facility) for help with mass spectrometry, and acknowledges the European Research Council (268893), Science Foundation Ireland (13/IA/1910) and the Wellcome Trust. Funding from the National Science Foundation (NSF DEB-0072741 and 0417180) supported discovery and study of many new yeast strains that contributed to this study. Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U.S. Department of Agriculture.

References

1. Dujon, B. Yeast evolutionary genomics. *Nat Rev Genet* **11**, 512-524 (2010).
2. Kurtzman CP, F.J., Boekhout T, ed. The Yeasts, a Taxonomic Study. (Elsevier, Amsterdam; 2011).
3. Nagy, L.G. et al. Latent homology and convergent regulatory evolution underlies the repeated emergence of yeasts. *Nat Commun* **5**, 4471 (2014).
4. Sylvester, K. et al. Temperature and host preferences drive the diversification of Saccharomyces and other yeasts: a survey and the discovery of eight new yeast species. *FEMS Yeast Res* **15** (2015).
5. Steensels, J. et al. Improving industrial yeast strains: exploiting natural and artificial diversity. *FEMS Microbiol Rev* **38**, 947-995 (2014).
6. de Queiroz, A. & Gatesy, J. The supermatrix approach to systematics. *Trends Ecol Evol* **22**, 34-41 (2007).
7. DeGiorgio, M. & Degnan, J.H. Fast and consistent estimation of species trees using supermatrix rooted triples. *Mol Biol Evol* **27**, 552-569 (2010).
8. Kawaguchi, Y., Honda, H., Taniguchi-Morimura, J. & Iwasaki, S. The codon CUG is read as serine in an asporogenic yeast Candida cylindracea. *Nature* **341**, 164-166 (1989).
9. Miranda, I., Silva, R. & Santos, M.A. Evolution of the genetic code in yeasts. *Yeast* **23**, 203-213 (2006).
10. Ohama, T. et al. Non-universal decoding of the leucine codon CUG in several Candida species. *Nucleic Acids Res* **21**, 4039-4045 (1993).
11. Osawa, S. & Jukes, T.H. On codon reassignment. *J Mol Evol* **41**, 247-249 (1995).
12. Schultz, D.W. & Yarus, M. On malleability in the genetic code. *J Mol Evol* **42**, 597-601 (1996).
13. Sugita, T. & Nakase, T. Non-universal usage of the leucine CUG codon and the molecular phylogeny of the genus Candida. *Syst Appl Microbiol* **22**, 79-86 (1999).
14. Santos, M.A., Gomes, A.C., Santos, M.C., Carreto, L.C. & Moura, G.R. The genetic code of the fungal CTG clade. *C R Biol* **334**, 607-611 (2011).
15. Santos, M.A., Ueda, T., Watanabe, K. & Tuite, M.F. The non-standard genetic code of Candida spp.: an evolving genetic code or a novel mechanism for adaptation? *Mol Microbiol* **26**, 423-431 (1997).

16. Muhlhausen, S. & Kollmar, M. Molecular phylogeny of sequenced Saccharomycetes reveals polyphyly of the alternative yeast codon usage. *Genome Biol Evol* **6**, 3222-3237 (2014).
17. Muhlhausen, S. & Kollmar, M. Predicting the fungal CUG codon translation with Bagheera. *BMC Genomics* **15**, 411 (2014).
18. Claudel-Renard, C., Chevalet, C., Faraut, T. & Kahn, D. Enzyme-specific profiles for genome annotation: PRIAM. *Nucleic Acids Res* **31**, 6633-6639 (2003).
19. Jeffries, T.W. et al. Genome sequence of the lignocellulose-bioconverting and xylose-fermenting yeast *Pichia stipitis*. *Nature Biotechnology* **25**, 319-326 (2007).
20. Wohlbach, D.J. et al. Comparative genomics of xylose-fermenting fungi for enhanced biofuel production. *Proc Natl Acad Sci U S A* **108**, 13212-13217 (2011).
21. Douglas, H.C. & Hawthorne, D.C. Regulation of genes controlling synthesis of the galactose pathway enzymes in yeast. *Genetics* **54**, 911-916 (1966).
22. Hittinger, C.T. & Carroll, S.B. Gene duplication and the adaptive evolution of a classic genetic switch. *Nature* **449**, 677-681 (2007).
23. Hittinger, C.T. et al. Remarkably ancient balanced polymorphisms in a multi-locus gene network. *Nature* **464**, 54-58 (2010).
24. Hittinger, C.T., Rokas, A. & Carroll, S.B. Parallel inactivation of multiple GAL pathway genes and ecological diversification in yeasts. *Proc Natl Acad Sci U S A* **101**, 14144-14149 (2004).
25. Slot, J.C. & Rokas, A. Multiple GAL pathway gene clusters evolved independently and by different mechanisms in fungi. *Proc Natl Acad Sci U S A* **107**, 10136-10141 (2010).
26. Webster, T.D. & Dickson, R.C. The organization and transcription of the galactose gene cluster of Kluyveromyces lactis. *Nucleic Acids Res* **16**, 8011-8028 (1988).
27. Wolfe, K.H. et al. Clade- and species-specific features of genome evolution in the Saccharomycetaceae. *FEMS Yeast Res* **15**, fov035 (2015).
28. Johnston, M. A model fungal gene regulatory mechanism: the GAL genes of Saccharomyces cerevisiae. *Microbiol Rev* **51**, 458-476 (1987).
29. Wegner, G.H. Emerging applications of the methylotrophic yeasts. *FEMS Microbiol Rev* **7**, 279-283 (1990).

30. Koivistoinen, O.M. et al. Characterisation of the gene cluster for l-rhamnose catabolism in the yeast Scheffersomyces (Pichia) stipitis. *Gene* **492**, 177-185 (2012).
31. Gojkovic, Z. et al. Horizontal gene transfer promoted evolution of the ability to propagate under anaerobic conditions in yeasts. *Mol. Genet. Genomics* **271**, 387-393 (2004).
32. Hall, C. & Dietrich, F.S. The reacquisition of biotin prototrophy in *Saccharomyces cerevisiae* involved horizontal gene transfer, gene duplication and gene clustering. *Genetics* **177**, 2293-2307 (2007).
33. Hanson, S.J., Byrne, K.P. & Wolfe, K.H. Mating-type switching by chromosomal inversion in methylotrophic yeasts suggests an origin for the three-locus Saccharomyces cerevisiae system. *Proc Natl Acad Sci U S A* **111**, E4851-4858 (2014).
34. Maekawa, H. & Kaneko, Y. Inversion of the chromosomal region between two mating type loci switches the mating type in Hansenula polymorpha. *PLoS Genet* **10**, e1004796 (2014).
35. Hickman, M.A., Froyd, C.A. & Rusche, L.N. Reinventing heterochromatin in budding yeasts: Sir2 and the origin recognition complex take center stage. *Eukaryot. Cell* **10**, 1183-1192 (2011).
36. Kurtzman, C.P. & Robnett, C.J. Relationships among genera of the Saccharomycotina (Ascomycota) from multigene phylogenetic analysis of type species. *FEMS Yeast Res* **13**, 23-33 (2013).
37. Nagy, L.G. et al. Latent homology and convergent regulatory evolution underlies the repeated emergence of yeasts. *Nature Communications* **5**, 8 (2014).
38. Jeffries, T.W. & Van Vleet, J.R. Pichia stipitis genomics, transcriptomics, and gene clusters. *FEMS Yeast Res* **9**, 793-807 (2009).
39. Liu, X. (2012) Conversion of the biodiesel by-product glycerol by the non-conventional yeast Pachysolen tannophilus. PhD thesis, Technical University of Denmark.
40. Wedlock, D.N. & Thornton, R.J. Transformation of a glucose negative mutant of Pachysolen tannophilus with a plasmid carrying the cloned hexokinase PII gene from Saccharomyces cerevisiae. *Biotetechnol Letters* **2**, 601-604 (1989).
41. Bezerra, A.R., Guimaraes, A.R. & Santos, M.A. Non-Standard Genetic Codes Define New Concepts for Protein Engineering. *Life (Basel)* **5**, 1610-1628 (2015).
42. Osawa, S. Evolution of the Genetic Code. (Oxford University Press, Oxford; 1995).

43. Henikoff, S. & Henikoff, J.G. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A* **89**, 10915-10919 (1992).
44. Matsuzawa, T. et al. New insights into galactose metabolism by Schizosaccharomyces pombe: isolation and characterization of a galactose-assimilating mutant. *J Biosci Bioeng* **111**, 158-166 (2011).
45. Suzuki, S., Matsuzawa, T., Nukigi, Y., Takegawa, K. & Tanaka, N. Characterization of two different types of UDP-glucose/-galactose 4-epimerase involved in galactosylation in fission yeast. *Microbiology* **156**, 708-718 (2010).